Data Analytics and Visualization

MIDTERM ESSAY

SP500 Index with Macroeconomics 2011 - 2023

*Presented by **Nguyen Quang Huy** - **Nguyen Gia Nhat Khanh***

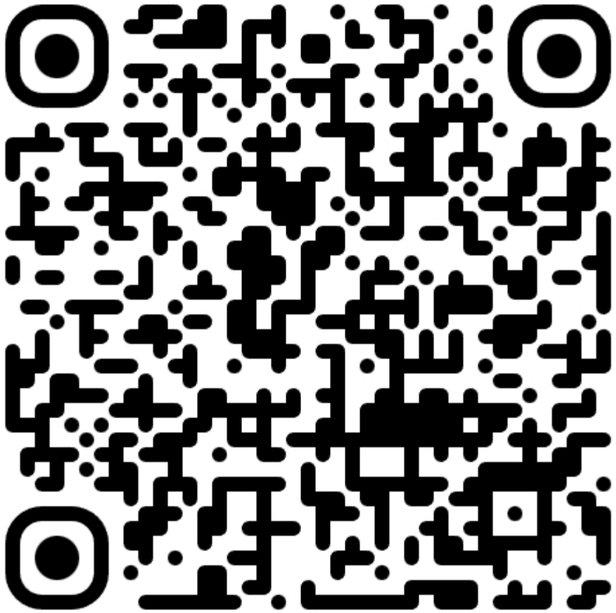
Faculty of Information Technology, Ton Duc Thang University

Email: {523h0140, 523h0149}@student.tdtu.edu.vn

QR Code

Presentation slide

Implementation



Github Resources

Outline

1. Introduction
2. Exploratory Data Analysis (EDA)
3. Probability Distribution Analysis
4. Hypothesis Testing
5. Correlation Analysis
6. Multiple Linear Regression
7. Conclusion

1. Introduction

Analyze the **S&P 500 Index** by applying **a structured data analysis** approach, using a **time series dataset** combining the S&P 500 Index with **macroeconomic indicators** (CPI, GLD, IEF, VIXY, UUP, USO) from *January 4, 2011 to December 12, 2023*.

2. Exploratory Data Analysis (EDA)

SP500 Daily index:

- From 1927 to 2024 with 24367 record of 5 variables

Variable	Description
Open	Opening price of the index on a given trading day
High	Highest price of the index during the trading day
Low	Lowest price of the index during the trading day
Close	Closing price of the index at the end of the trading day
Volume	The total number of shares traded for the index constituents on that day

2. Exploratory Data Analysis (EDA)

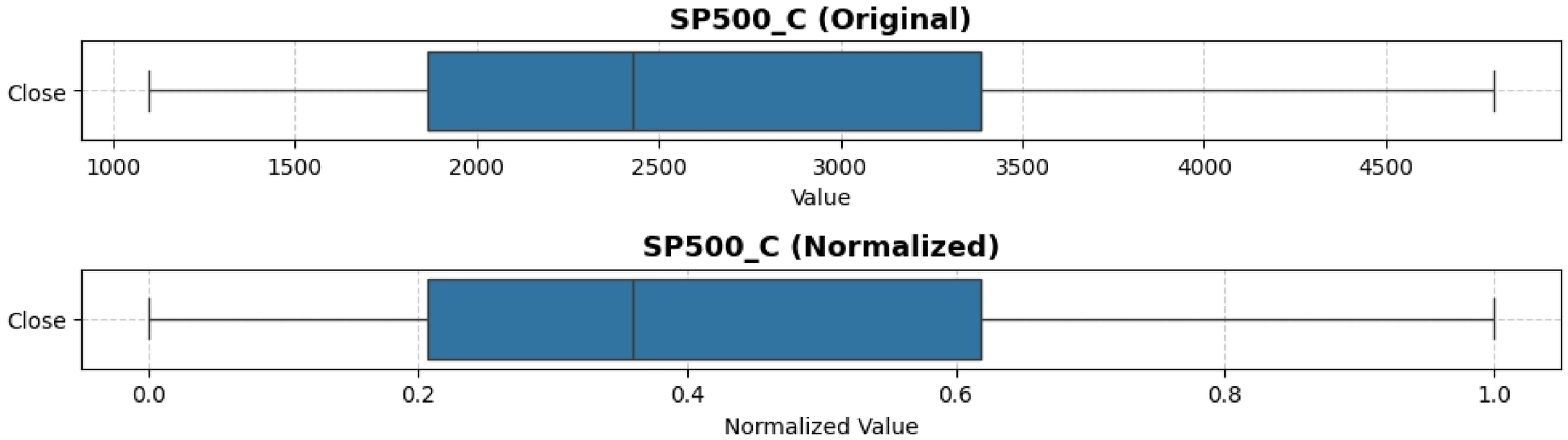
Macroeconomics Daily index:

- From 2000 to 2024 with 6000 record of 6 variables

Variable	Description
VIXY	CBOE Volatility Index, representing expected market volatility
IEF	U.S. 10-Year Treasury Yield, reflecting long-term interest rate expectations
UUP	Measuring the value of the U.S. dollar against a basket of major currencies
USO	Crude oil price per barrel
GLD	Gold spot price per ounce, representing a key safe-haven asset
CPI	U.S. Consumer Price Index, representing inflation levels

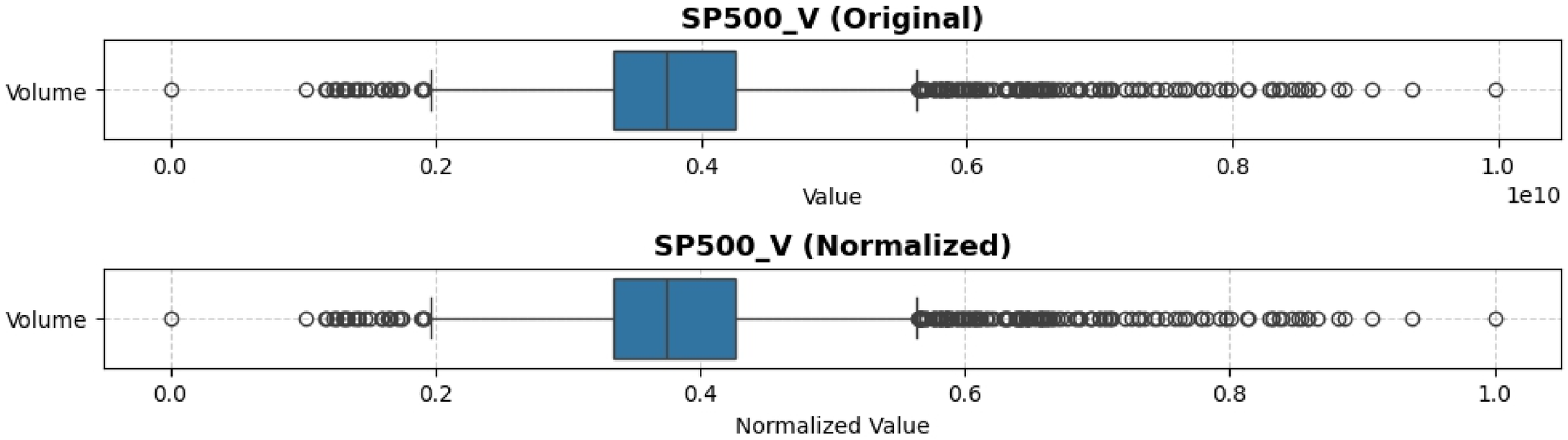
2. Exploratory Data Analysis (EDA)

Close Index



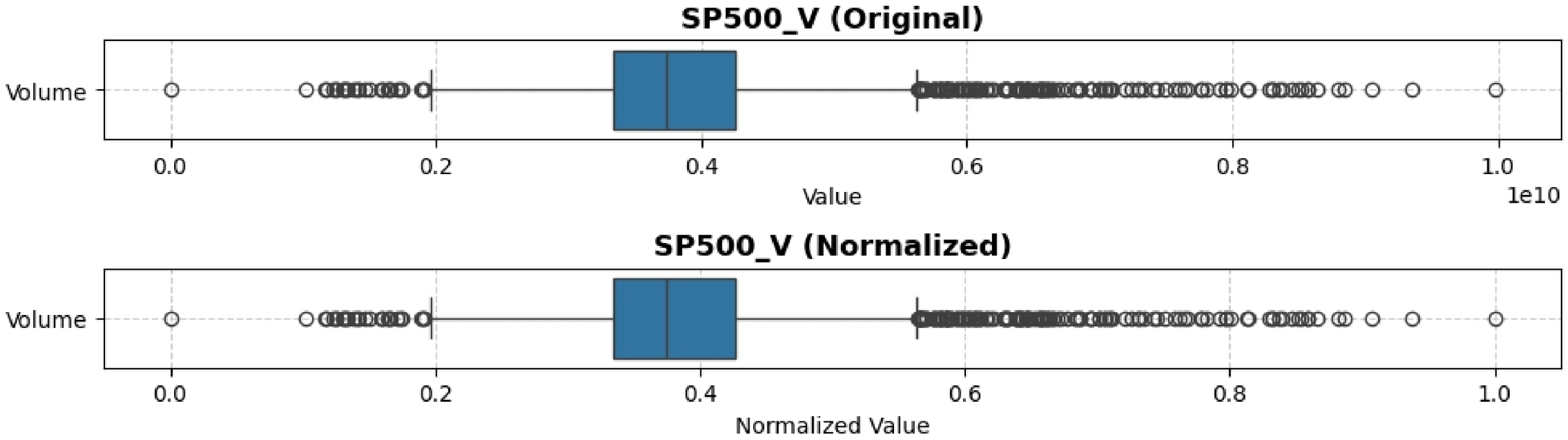
2. Exploratory Data Analysis (EDA)

Trading Volume



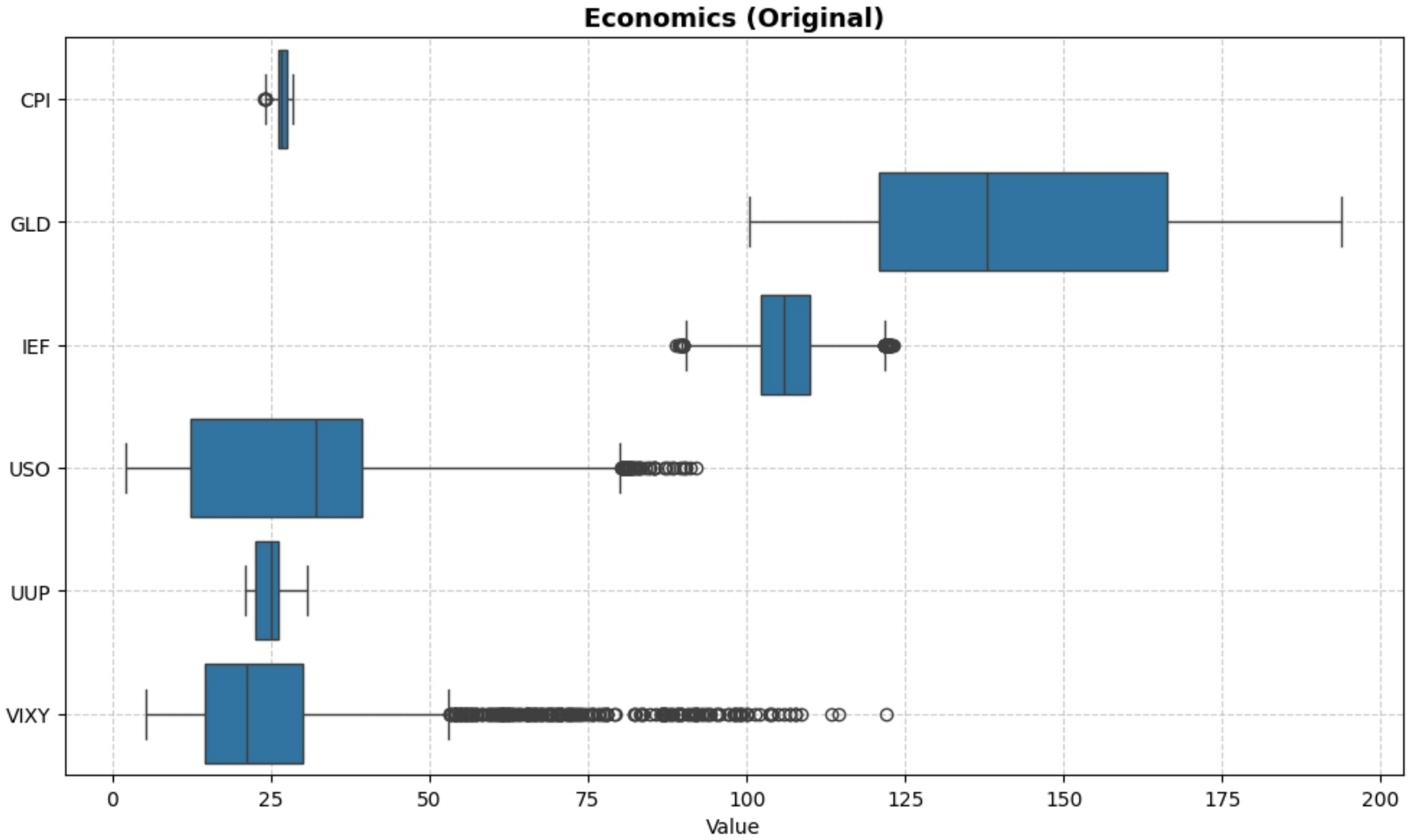
2. Exploratory Data Analysis (EDA)

Trading Volume



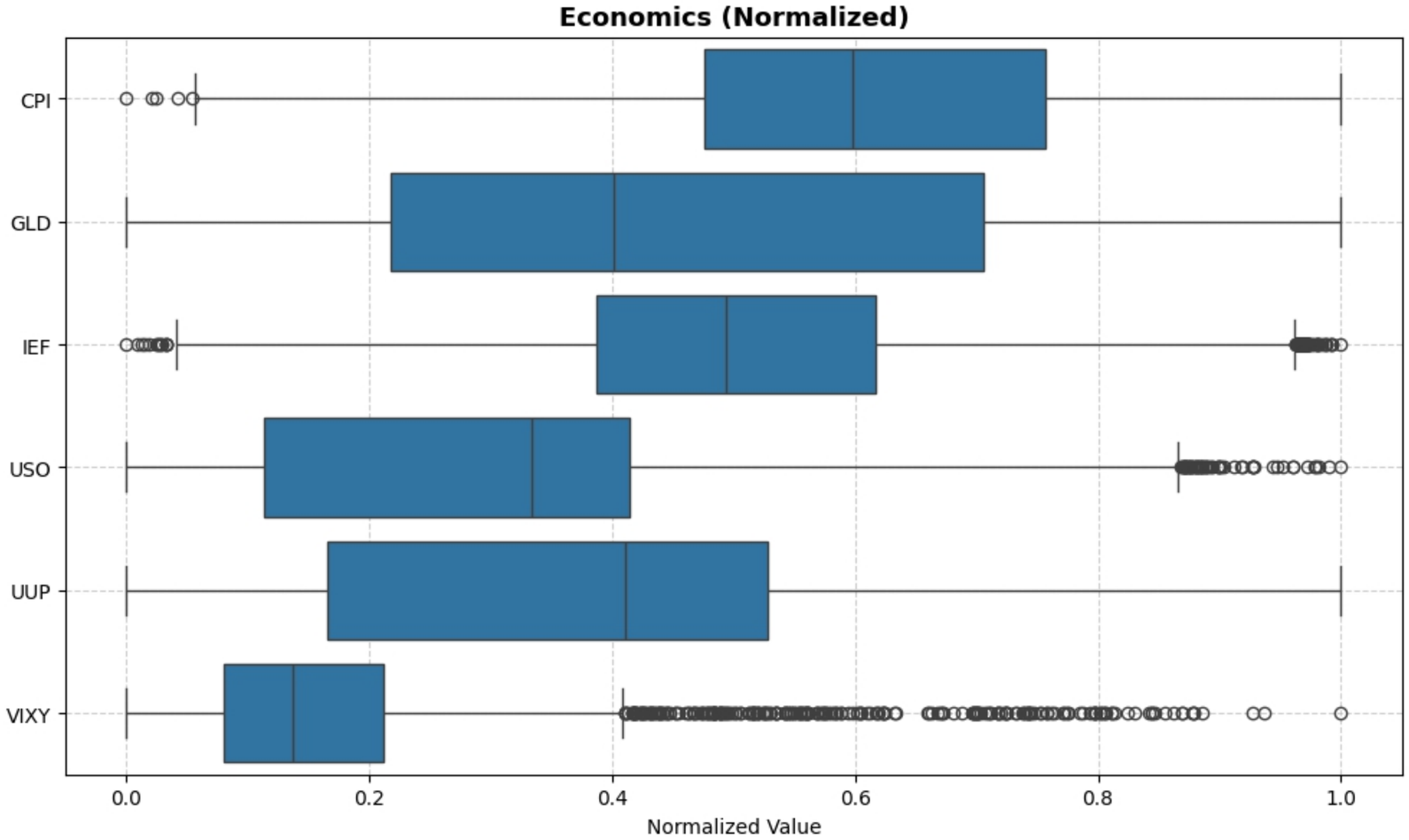
2. Exploratory Data Analysis (EDA)

Macro economics



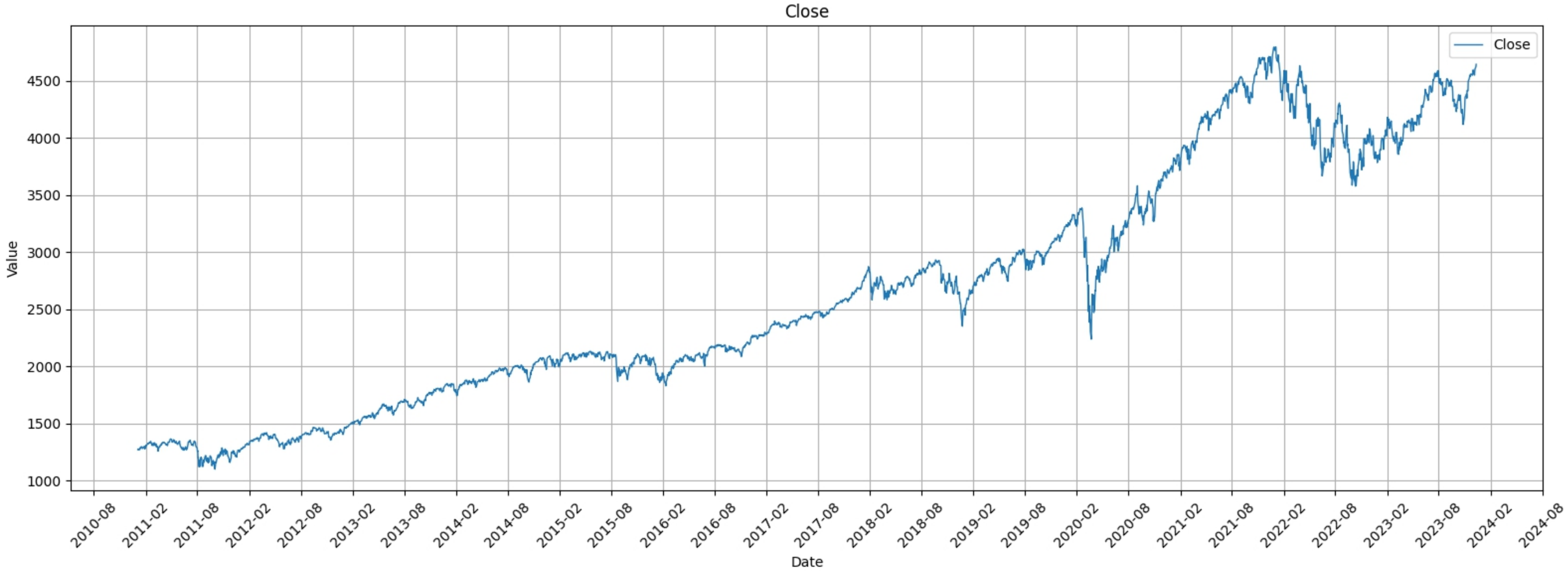
2. Exploratory Data Analysis (EDA)

Macro economics (Normalized)



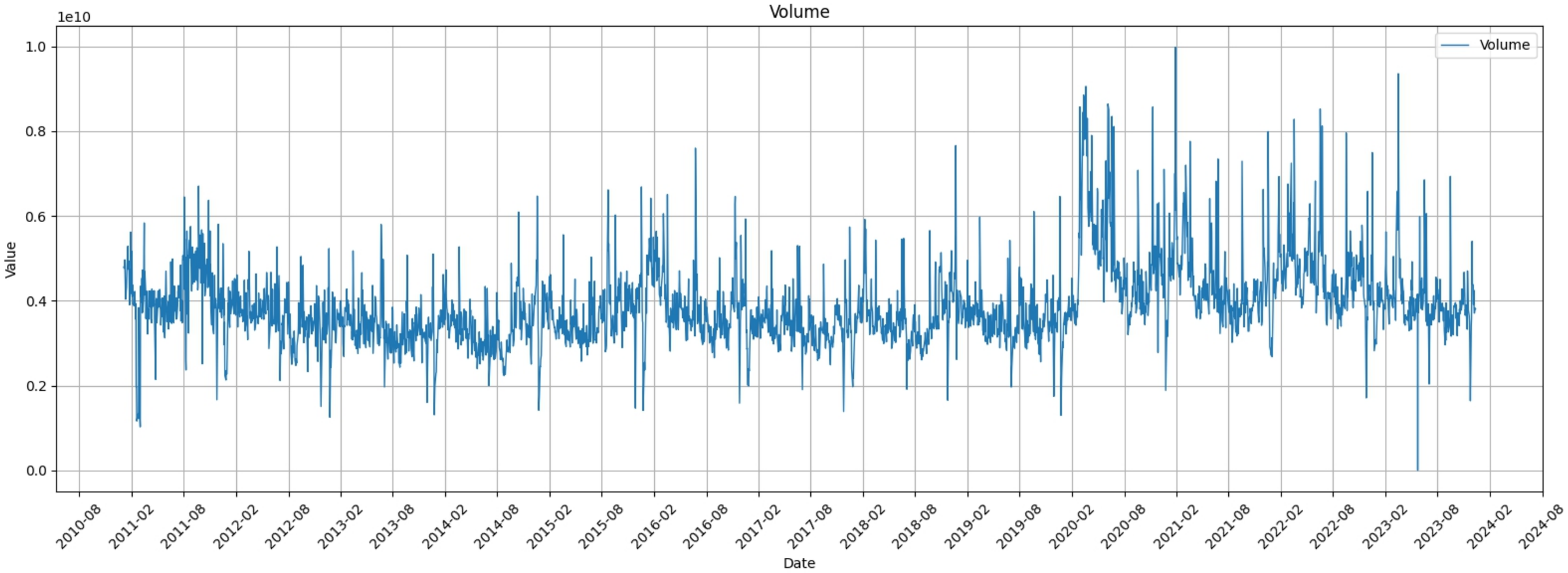
2. Exploratory Data Analysis (EDA)

Line Chart (Close Index)



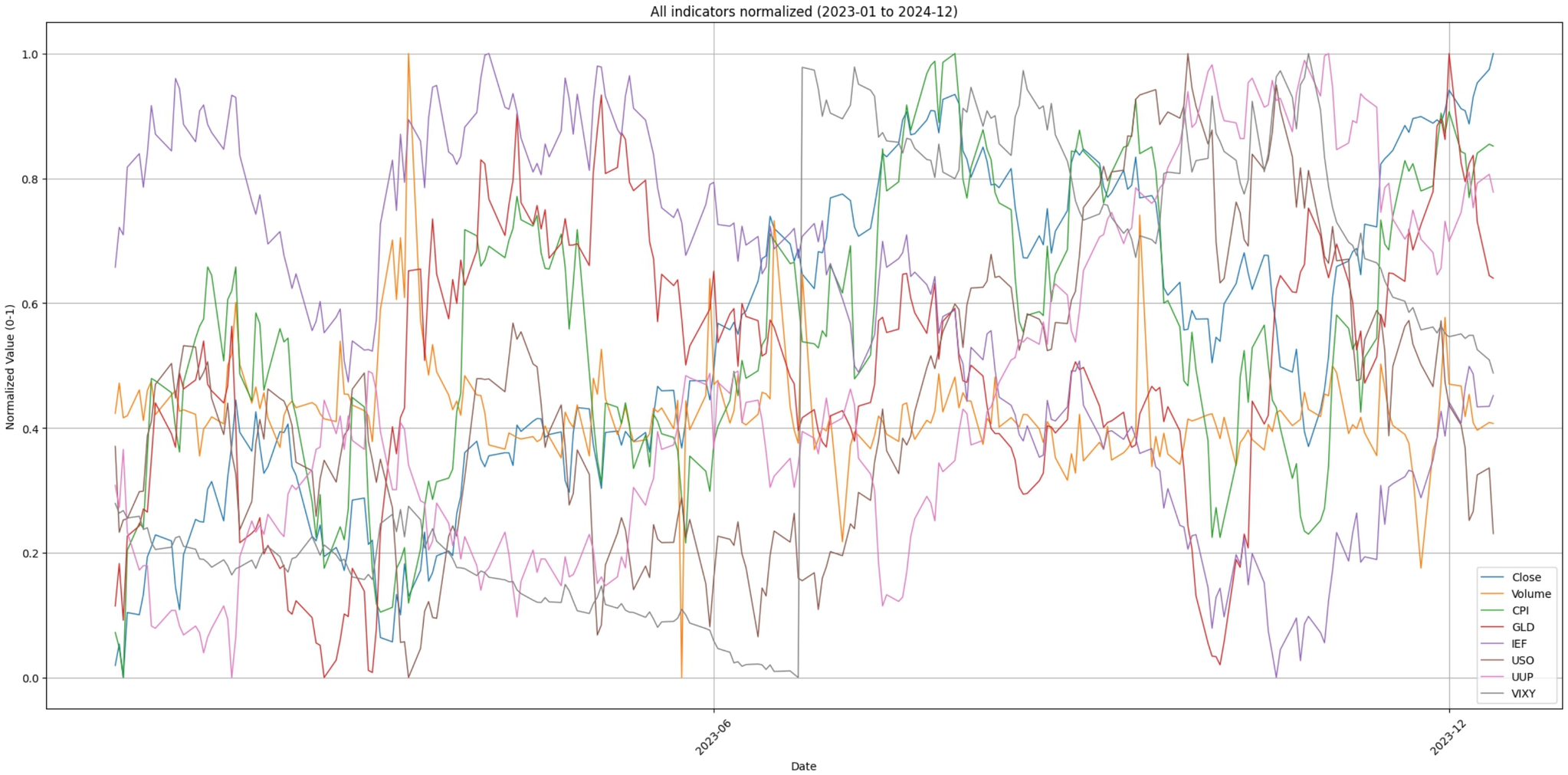
2. Exploratory Data Analysis (EDA)

Line Chart (Trading Volume)



2. Exploratory Data Analysis (EDA)

Line Chart (Combined)



2. Exploratory Data Analysis (EDA)

Outlier Detection

1. Z-score: Identifies outliers by measuring how many standard deviations a point is from the mean.

2. IQR: Detects outliers as values falling below Q1 − 1.5×IQR or above Q3 + 1.5×IQR.

3. Rolling Z-score: Applies Z-score on a moving window to find outliers in time series data.

4. Seasonal Decomposition Residuals: Flags outliers as large deviations in the residual component after decomposing trend and seasonality.

5. Isolation Forest: Uses an ensemble of random trees to isolate anomalies that require fewer splits to separate.

2. Exploratory Data Analysis (EDA)

Outlier Detection

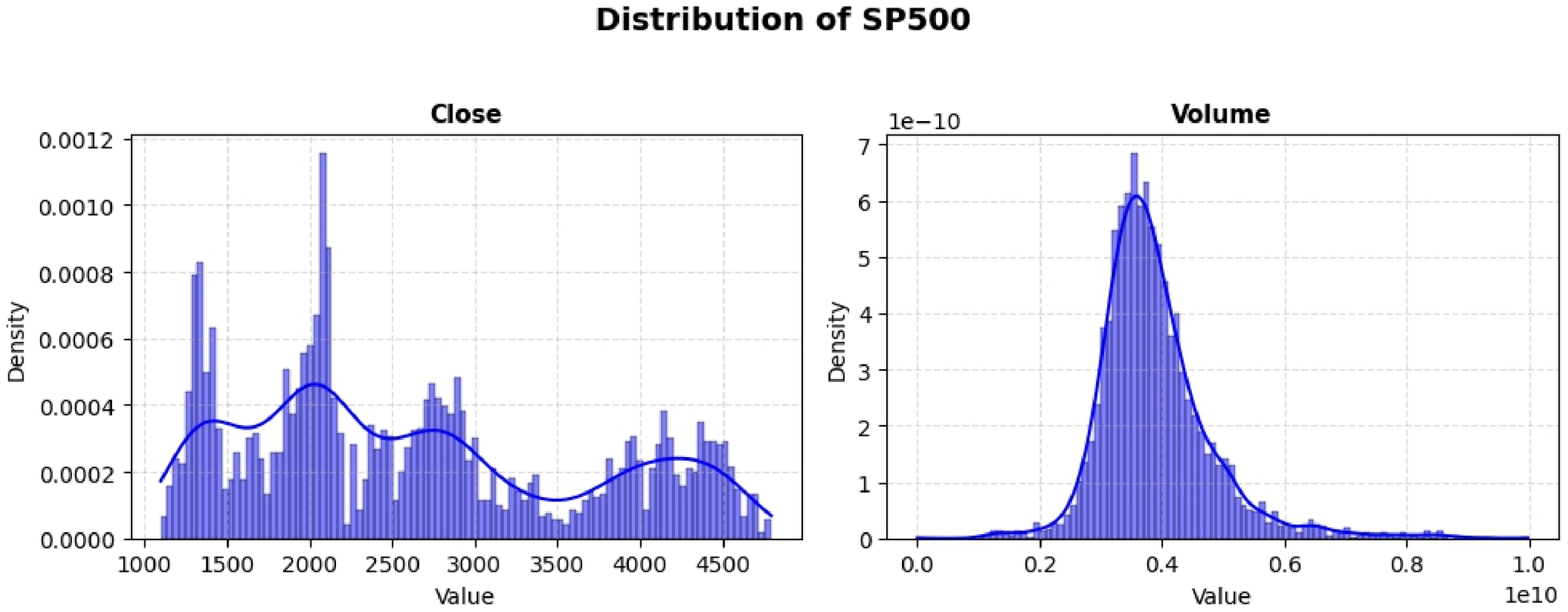
Date	2020-03-26	2020-03-27	2020-03-28	2020-03-28
z_close	True	True	False	False
iqr_close	False	False	True	False
decomp_close	False	True	False	False
arima_close	False	False	False	False
iso_multivariate	False	True	True	False
rolling_close	False	False	True	False
any_outlier	True	True	True	False
n_methods_flagged	1	3	3	0

2. Exploratory Data Analysis (EDA)

Outlier Detection

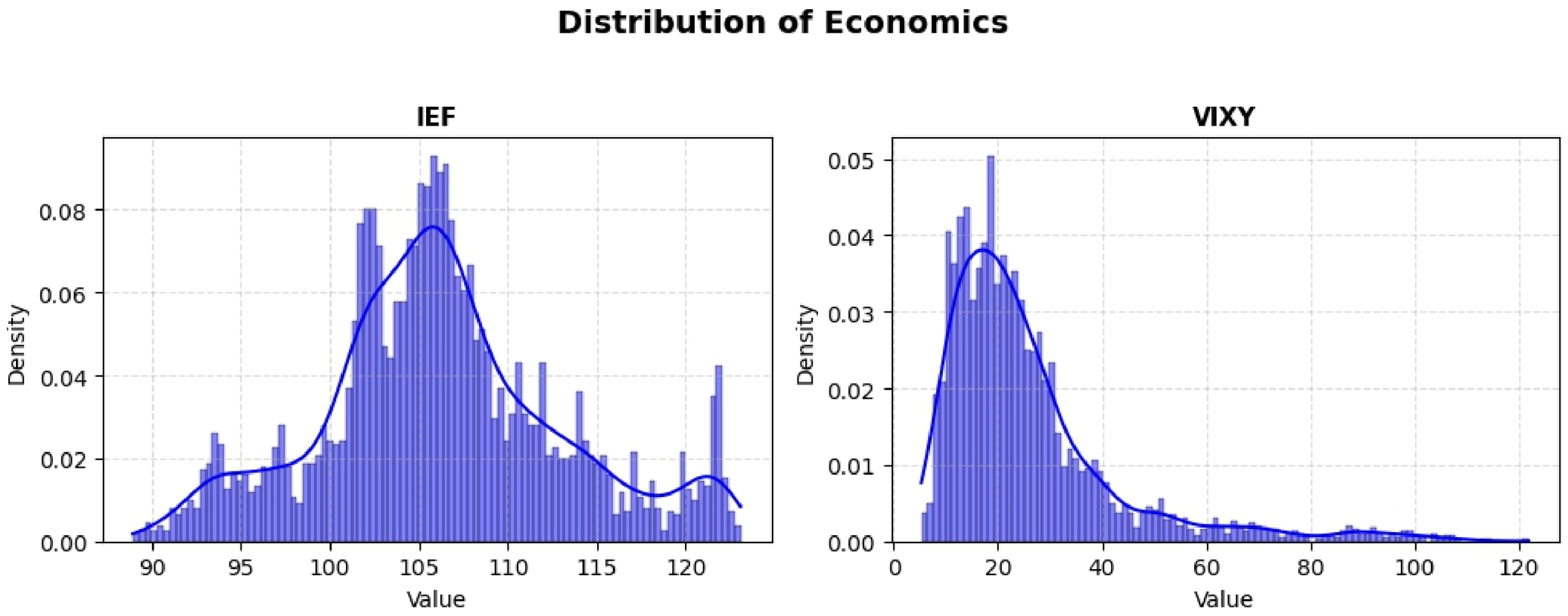
3. Probability Distribution Analysis

Close Index and Trading Volume



3. Probability Distribution Analysis

VIXY and IEF



3. Probability Distribution Analysis

Normality Test and Distribution Fit

Apply Shapiro-Wilk and Kolmogorov-Smirnow test

Index	Variable	Shapiro_p	KS_p	Distribution
0	Close	4.426160e-37	3.073666e-43	Non-normal
1	Volume	2.883589e-42	2.667348e-31	Non-normal
2	IEF	2.947230e-21	5.636244e-14	Non-normal
3	VIXY	1.150930e-54	5.995955e-74	Non-normal

4. Hypothesis Testing

T-test

Is there a significant difference in trading volume between periods of low volatility and high volatility in the stock market?

- Null hypothesis: The mean trading volume during low-volatility and high-volatility periods are equal.
- Alternative hypothesis: The mean trading volumes differ between the two periods.

4. Hypothesis Testing

T-test

The dataset was divided into two groups based on the median of VIXY:

- Low-volatility period: VIXY < median(VIXY)
- High-volatility period: VIXY ≥ median(VIXY)

Statistics	Value
t-staitstics	2.325
p-value	0.02012

Answer: There is a statistically significant difference in trading volume between low- and high-volatility periods

4. Hypothesis Testing

ANOVA test

Is there a significant difference in the S&P 500 closing price (Close) among different economic periods?

- Null hypothesis: The mean closing prices are the same across all economic periods.
- Alternative hypothesis: At least one period has a different mean closing price.

The data was divided into three economic phases:

- Before COVID-19: 2011–2019
- During COVID-19: 2020–2021
- After COVID-19: 2022–2023

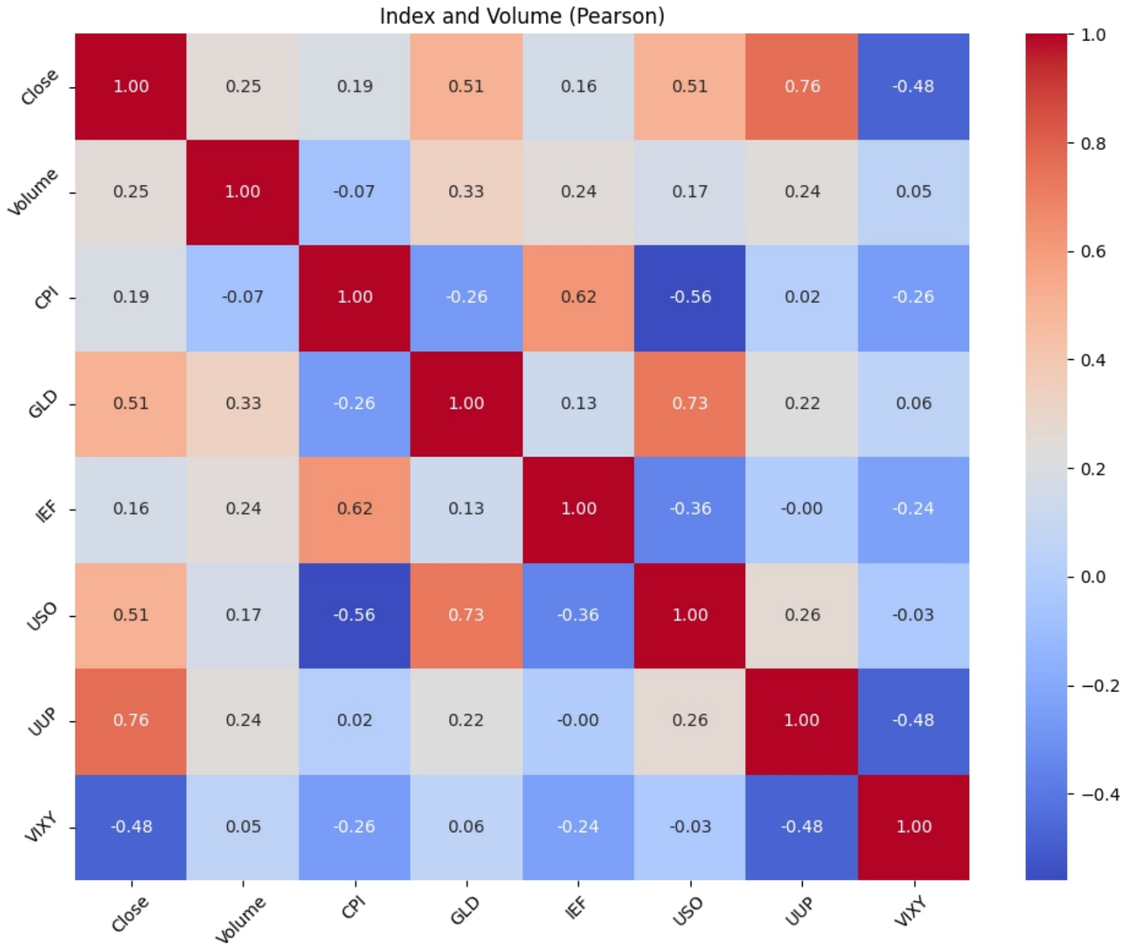
4. Hypothesis Testing

ANOVA test

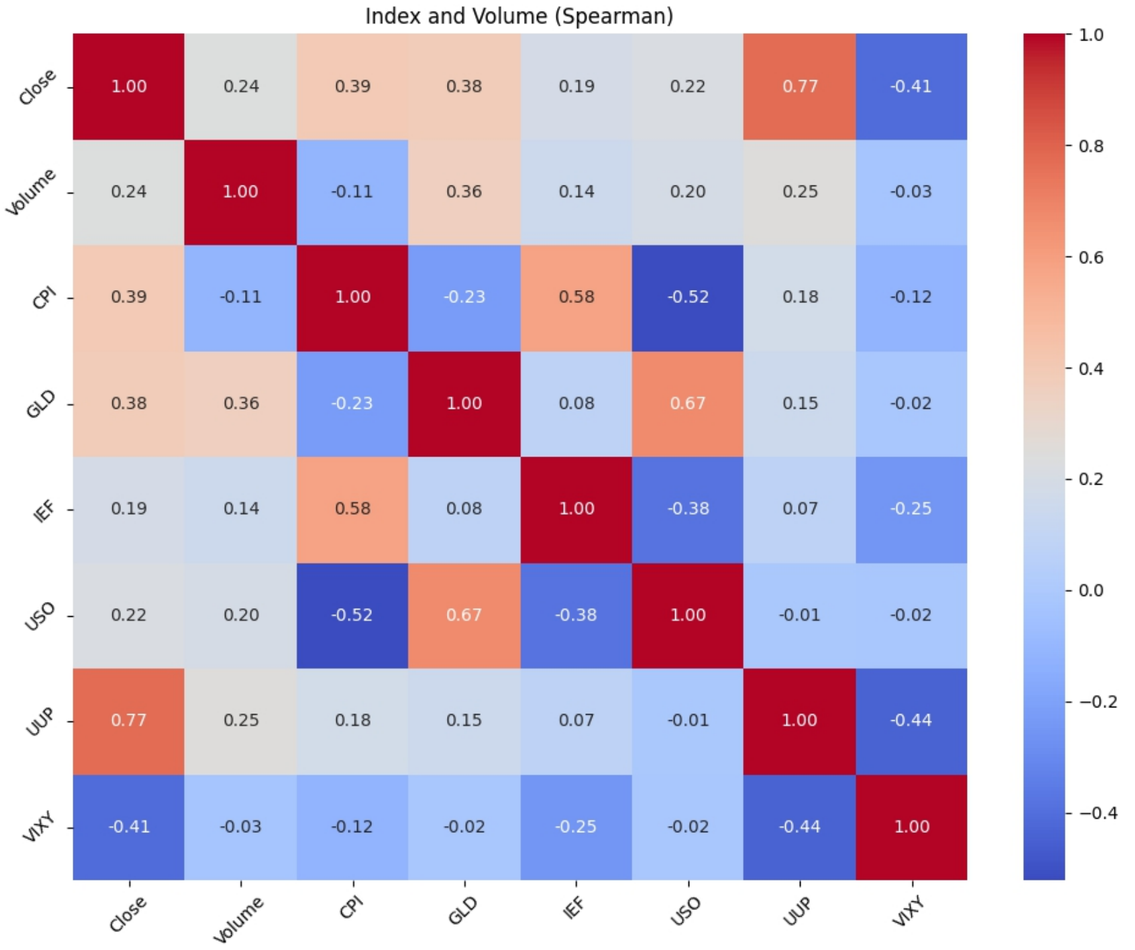
Statistics	Value
f-staitstics	2.325
p-value	0.02012

Answer: Therefore, there are significant differences in the mean closing price among three economic periods.

5. Correlation Analysis



Pearson



Spearman

6. Multiple Linear Regression

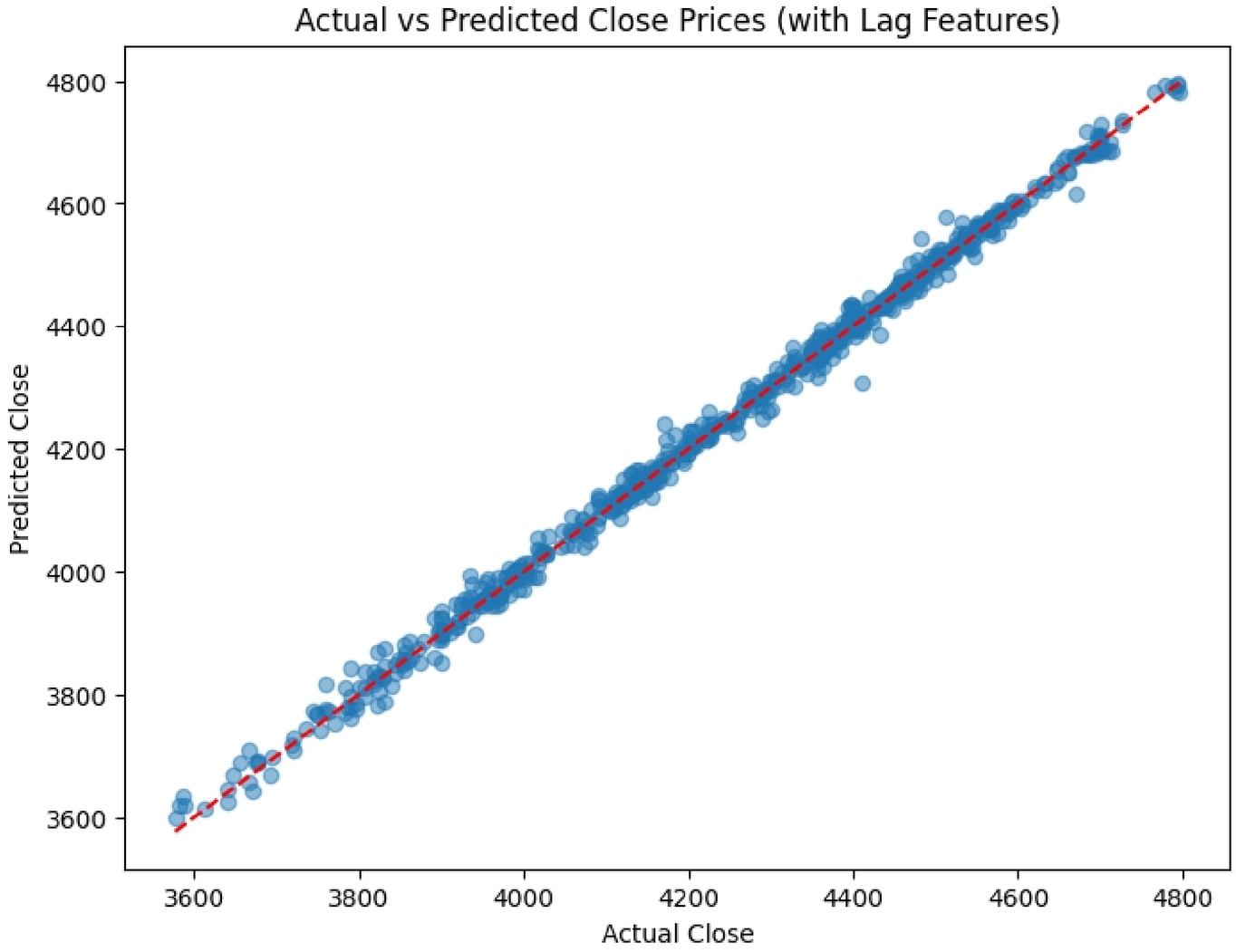
Model Implementation

- Target (y): Close
- Predictors (X): Open, High, Low, Volume, CPI, GLD, IEF, USO, UUP, VIXY, Close_lag1, Close_lag2, Close_lag3
- The model was trained using Linear Regression with a time-based train-test split (80%-20%).

6. Multiple Linear Regression

Model Evaluation

Metric	Value
R^2	0.9962
RMSE	16.9717



References